同時通訳品質評価方法検討のための同時通訳者と翻訳者の評価比較分析

蒔苗 茉那[1] 須藤 克仁[1] 中村 哲[1] 松下 佳世[2] 山田 優[2]
[1]奈良先端科学技術大学 [2]立教大学
{makinae.mana.mh2, sudoh, s-nakamura}@is.naist.jp

概要

同時通訳の品質評価は，人と機械双方の同時通訳の向上と発展のために重要である．しかし，標準的な品質評価基準はまだ確立されていない．本研究では，同時通訳の品質評価基準を検討するための第一歩として，同時通訳者と翻訳者による評価の比較分析を行い，通訳と翻訳の品質評価観点の差異について議論する．評価は同時通訳者と翻訳者各 1 名が，同一の記者会見に対して Multidimensional Quality Metrics (MQM) をもとにした「JTF 翻訳品質評価ガイドライン」を利用した． 評価結果の比較分析により，同時通訳者と翻訳者の間の評価観点の共通点と相違点を明らかにする．そして同時通訳品質評価への応用の可能性を検討し、改良が必要な項目を明らかにする．

1 はじめに

同時通訳とは，原言語から目的言語へ，発話された内容を聞き取りつつ，同時に発話内容を訳出していく行為を指す．同時通訳では， 訳出は原発話になるべく追いつくことが要求されるため，厳しい時間制約の下で話者の伝えたいことを正確に訳出することが求められる．

しかし，そのような要求に応えられているかの標準的な評価基準は確立されていない．同時通訳の訳出結果が原発話の内容を十分汲み取れているかを評価するためには、訳出結果の品質の良し悪しを多角的に判断しなければならないが、そのための具体的かつ実用的な基準はまだ存在しないのである．ここでの評価は通訳品質の向上につながることから，現在人間が行っている同時通訳を評価する上でも有益である．しかし現状，同時通訳の訳出結果全体を大まかに評価する指標はあるものの，一文ずつなど細かいセグメント単位での訳出に注目した評価指標はなく，十分な評価ができていない状況にある．

一方で近年，機械による自動同時翻訳の研究が進んでいる．例えば Fukuda ら[1]は漸進的な音声認識，機械翻訳，音声合成をカスケード処理する自動同時翻訳システム実現した．機械もまた人間と同様に，訳出内容を評価することができれば，その評価をもとに改善することができるため，評価は重要である．しかし，同時通訳の品質評価基準が存在しないことから， 現状は翻訳の指標である BLEU 等を元に改善を試みているが，これらは評価指標に不十分である．なぜなら，通訳評価には流暢性，自然性より，同時性，内容伝達度の視点が重要となるためである．なぜなら，自動同時通訳の改善に翻訳の指標が用いられているということは，同時通訳の特徴が改善指標に十分に反映されていないことを意味するからである．そのため，同時通訳の品質評価基準を構築することは，人間だけでなく機械による同時通訳にとっても有益であると考える．

本研究では，同時通訳の品質評価基準構築に向けて，翻訳の品質評価基準である Multidimensional Quality Metrics (MQM) [2]に着目し，その同時通訳評価への応用について検討する．原言語から目的言語に変換するという点で同時通訳と翻訳は似ているものの，一方で訳出時間の制限などいくつかの点で異なる性質を持つことから，翻訳の品質評価基準を同時通訳にそのまま当てはめようとすると問題が生じる．例えば，聞き手の内容理解に与える影響が小さいと考えられる程度の簡略化や省略は、時間的な制約を伴う同時通訳においては、一定程度許容されるからである．そこで本稿ではそのための第一歩として，MQM を同時通訳評価へ応用するにあたり何が必要かを明らかにするために，同時通訳者と翻訳者による評価の視点の共通点と相違点の分析を行う．具体的には，同一の同時通訳に対して同時通訳

者と翻訳者が MQM に基づく JTF 翻訳品質評価ガイドライン[3]を利用して行った評価結果の比較と分析を通じて，翻訳評価の視点をほぼそのまま応用可能な項目および同時通訳評価向けに改変が必要な項目を明らかにすることを目指す．

本稿の検討を通じ以下が明らかになった。

- MQM 評価項目のうち，同時通訳評価に応用可能な項目はカテゴリ「正確さ-誤訳-」である．
- MQM 評価項目のうち，同時通訳評価に応用するには要検討な項目は流暢さ」とカテゴリ「抜けと余分」である．

2 関連研究

MQM（Multidimentional Quaity Metrics）とは，翻訳の評価指標フレームワークであり，この MQM を用いた翻訳の評価手法の有効性が示されている．MQM については 3 章で詳説する．Freitag ら[4]は，Direct Assessment による人手評価よりも，MQM 評価が Scalar Quality Metric (SQM)など他の評価手法と相関が最も強いことを示し，機械翻訳の評価に MQM を用いるのが適当ではないかと提案している．WMT21 Metrics Shared Task [5] では、実際に MQM を用いた人手評価が追加で行われており，MQM の評価手法としての有効性を示している．

既存の同時通訳の評価指標は，訳出結果全体を通じたプロダクト評価というよりも，むしろ通訳者を評価するものであり，代表的なものに NAATI Metrics [6]と EU Metrics [7]がある．NAATI (National Accreditation Authority For Translators and Interpreters）とは，オーストラリア政府公認の翻訳者や通訳者の国家資格の認定のことを指し，そこで使用されている NAATI Metrics は通訳の内容を A: Meaning Transfer Skill， B: Application of interpreting mode， C: Rhetorical skill， D: Language Proficiency enabling meaning transfer の 4 つの指標をもとに 5 段階評価を行うことで，通訳全体を通して適切な訳出が行われていたかを評価する．EU Metrics では，Content， Delivery/From， Technique の 3 つの指標をもとに，通訳全体を通して適切な訳出が行われていたか評価をする．一方で，通訳者ではなく，通訳内容に注目し，一文ずつなど細かいセグメント単位での通訳の評価指標はこれまでのところ確認できていない．

3 MQM による同時通訳評価

3.1 MQM・JTF 翻訳品質評価ガイドライン

MQM とは翻訳の評価指標フレームワークであり，日本翻訳連盟（JTF）では MQM をもとに JTF 翻訳品質評価ガイドライン[3]を作成，公開している．

MQM や JTF 翻訳品質評価ガイドラインでは，評価依頼者が各エラーカテゴリの重みを事前に設定することで仕様を決定し，評価者はその仕様をもとに，翻訳結果を 1 文単位でエラーベースの評価を行う．具体的には重みと重大度を掛け合わせ点数化する．エラーのカテゴリには正確さ・流暢さ・用語・スタイル・地域慣習などがある．評価依頼者が各エラーカテゴリに対し重みを自由に設定するので，どのエラー項目を重視度するかの表現が可能である。エラーが文章に与える重大度の程度は深刻・重度・軽度・なしがあり，JTF 翻訳品質評価ガイドラインのサンプル評価シートにおけるそれぞれの標準値は 100・10・1・0 であり，重大なエラーには大きなペナルティがかかるようになっている。この重み付き和で表されるエラースコアが大きいほど、その文章に対するエラーは深刻であると考える．

3.2 同時通訳評価への応用

本研究では，MQM に基づく同時通訳評価の確立に向けた検討を行う．過去の同時翻訳評価の試み[8]では，英日の同時翻訳タスクに対して，JTF 翻訳品質ガイドラインを用いて，翻訳者が評価者として訳出内容を評価した．その結果，BLEU スコアと人手評価によるエラースコアとの間に正の相関関係が見られたことが報告されている．しかしながら先に述べたような翻訳と通訳の違いについては十分に考慮されておらず，通訳においては問題とされないような点について過剰な減点を受けていることが憂慮される．そこで本研究では通訳者と通訳者による評価を実施し，評価結果の比較と分析を行う．

今回の分析から期待される効果として，両者間での共通点が明らかになれば，それは翻訳内容の品質を評価する MQM は同時通訳の品質評価にも応用可能であると示すことができる．また両者間での相違点が明らかになれば，それは同時通訳評価者と翻訳評価者の間で評価すべき項目の違いが存在することの現れであり，それらの違いは同時通訳評価体系の構築時に検討すべき項目と考えることができる．

分析のために，同時通訳者，翻訳者，各 1 名に，共通の同時通訳データの評価を依頼した．同時通訳者は 15 年以上の実務経験を持つが、JTF ガイドラインを使用した評価の経験はない。翻訳者は JFT ガイドラインを使用した評価の経験を持つ。同時通訳データは日本記者クラブの記者会見の同時通訳収録データ(JNPC063) [9]である．評価にあたり同時通訳者には原発話と同時通訳の音声およびそれらの書き起こしを提示し，翻訳者には原発話と同時通訳の書き起こしのみを提示した．なお，評価を行った翻訳者は JTF ガイドラインを用いた評価作業の経験を有していたのに対し，同時通訳者は MQM や JTF ガイドラインに基づく評価作業の経験がなかったため，事前の打ち合わせにより JTF ガイドラインに基づく評価方法について説明を行った上で，エラー重大度について通訳の観点で評価して良い旨を伝えた．

表 1　JTF ガイドラインによる同時通訳者と翻訳者の同時通訳評価結果比較

	カテゴリ	原発話書き起こし	同時通訳書き起こし	同時通訳者評価（エラー値）	翻訳者評価（エラー値）
(1)	正確さ -誤訳-	it's for the government to make sure that the private sector will be able to take part in the economic progress of our country in the right way within the rule of law.	ですから政府としては民間がきちっと経済の進歩に適切な形で法の支配のもとで関与できること参加できることを願っていると思います。	the government to **make sure** …" →「政府としては願っていると思います」は誤訳。「政府として～を確実にする」が正しい（26.6）	「**make sure**」を「願っていると思います。」と訳すのは不適切（1.33）
(2)	流暢さ -英文の語順通りに訳されているので不自然-	It is now 30 years ago since I left Kyoto after a year which started in 1985 it ended in nineteen eighty six.	30 年ぶりですけれども 1985 年か 1986 年まで滞在したことがございます。	なし(0)	**英文の語順通りに訳出**されているため表現が不自然で流暢さに欠ける(0.67)
(3)	流暢さ -表現が不自然で流暢さに欠ける-	And you know that there have been attacks on police outposts and there was one attack on a police outpost as late as yesterday when one policeman was killed.	警察の攻撃ということも言われておりますけれども例えば昨日でも警察官が 1 人殺されてることがあります。	なし(0)	**表現が不自然**で流暢さに欠ける(0.67)
(4)	正確さ -抜けと余分-	And I am also confident that our relations with other countries would get better with time.	ただ他の国々との関係も時間が経つにつれさらに良くなっていくだろうと	なし(0)	「**And I am also confident that」部分が訳されていない**(1.33)

4 分析結果

以下に，JTF ガイドラインを用いた同時通訳者と翻訳者による同時通訳評価結果について，評価の共通点・相違点について分析した結果のうち代表的なものを例示する（表１）．表中のエラー値は翻訳者による評価において事前に設定したカテゴリおよび重大度による重みに基づくものである（重みの詳細は付録に記載）．

4.1 同時通訳者評価と翻訳者評価の共通点

同時通訳者と翻訳者による評価で共通する傾向があったものとして，例(1)に示すカテゴリ「**正確さ-誤訳-**」が挙げられる．当該カテゴリでは，原発話で用いられている単語が目的言語にて適切な訳に置き換わっていない場合は共通して減点されていた．エラー重大度の判定については同時通訳者と翻訳者の間で揺れがあり，例(1)では同時通訳者のほうが重大度が高いとしており，エラー値が 26.6 となったのに対し，翻訳者では 1.33 であった．

4.2 同時通訳者評価と翻訳者評価の相違点

同時通訳評価者と翻訳評価者の間における評価項目の相違点として，**カテゴリ「流暢さ」**と**カテゴリ「抜けと余分」**が挙げられる．カテゴリ「流暢さ」にて，翻訳評価者は目的言語が英文の語順通りに訳されているので不自然（例(2) エラー値 0.67），表現が不自然で流暢さに欠ける（例(3) エラー値 0.67）という理由で，それら理由に該当する訳出文を減点対象としていた．一方同時通訳評価者は，前述した理由で訳出文を減点することはなかった．英日翻訳における英語と日本語の語順の違いから，同時通訳において早く訳出しようとすると通常の日本語の伝達順と入れ替えて訳出せざるを得ないが，翻訳の観点では自然さや流暢さが損なわれていると判断したからではないかと考えられる．例(2)であれば，「〜してから 30 年ぶりである」と時間の言及を強調する訳出を好ましい翻訳とすると，この通訳結果ではそうした強調が反映されていないと考えることもできる．

カテゴリ「抜けと余分」にて，翻訳評価者は原発話に含まれる単語が全て訳出されていない場合は基本的に減点としていた（例(4) エラー値 1.33）．一方同時通訳評価者は，原言語側に含まれる単語が全て訳出されていなかったとしても，全てを一律に減点することはなかった(エラー値 0)．こちらもカテゴリ「流暢さ」と同様に，英日の文法構造の違いと原言語になるべく追いつこうとする訳出時間制限のトレードオフから，同時通訳において，ある程度の抜けは容認されて然るべきなのではないかと考える．

5 考察

分析結果から，カテゴリ「正確さ-誤訳-」にて，同時通訳評価者と翻訳評価者の評価項目の視点が共通していることが分かった．そのため，カテゴリ「正確さ-誤訳-」については，翻訳内容の品質を評価する MQM・JTF ガイドラインに基づいて同時通訳の品質を評価することができると示唆される．しかしながら今回は評価対象，評価者とも事例が非常に限定されていることから，今後より多くの事例を通じて誤訳の認定基準について検討することが必要と考えられる．

一方，カテゴリ「流暢さ」とカテゴリ「抜けと余分」においては，同時通訳評価者と翻訳評価者の評価項目の視点が異なることが分かった．そのため，これらの違いは同時通訳の特徴と捉えることができ．こうした特徴を反映した評価体系の必要性が確認できた．

6 おわりに

本稿では，翻訳品質評価基準である Multidimensional Quality Metrics (MQM)が同時通訳評価へ応用可能であるかを調査するために，同時通訳評価者と翻訳評価者の間における評価の視点の共通点と相違点の分析を行い，応用可能な項目と検討が必要な項目を明らかにした。

一方で今回行った分析は，原言語と目的言語が一対一で評価される単文評価による分析であった．訳出時間や前後の文章の重要度によって訳出結果が大きく影響を受ける同時通訳において，前後の関係や訳出時間が考慮されない単文評価のみの分析結果は，同時通訳品質評価項目の検討に必要ではあるものの十分であるとは言えない．今後は，訳出時間と訳出結果，エラーの関係性について分析を進めていくことで，単文評価で分析するができなかった，同時通訳が持つ特徴であるリアルタイムで連続した文章間の関係性について明らかにしていきたいと考える．

謝辞

本研究は JSPS 科研費 JP21H05054 の助成を受けたものです.

参考文献

[1] Ryo Fukuda, Sashi Novitasari, Yui Oka, Yasumasa Kano, Yuki Yano, Yuka Ko, Hirotaka Tokuyama, Kosuke Doi, Tomoya Yanagita, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura. “Simultaneous Speech-to-speech Translation System with Transformer-based Incremental ASR, MT, and TTS,” Proc. Oriental COCOSDA 2021, 186-192, Nov. 18, 2021

[2] A.Lommel et al., “Multidimensional Quality Metrics (MQM):A Framework for Declaring and Describing Translation Quality Metrics,” Revista Tradumàtica: tecnologies de la traducció, 2014

[3] 一般社団法人　日本翻訳連盟（JTF），“JTF 翻訳品質評価ガイドライン”: https://www.jtf.jp/pdf/jtf_translation_quality_guidelines_v1.pdf

[4] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. “Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation,” Transactions of the Association for Computational Linguistics, 9:1460–1474. 2021

[5] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In Proceedings of the Sixth Conference on Machine Translation, pages 733–774, Online. Association for Computational Linguistics

[6] NAATI Metrics: https://www.naati.com.au/wp-content/uploads/2020/10/Certified-Interpreter-Assessment-Rubrics.pdf

[7] EU Mettics: https://europa.eu/interpretation/doc/marking_criteria_en.pdf

[8] Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, et al.. 2022. Findings of the IWSLT 2022 Evaluation Campaign. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 98–157, Dublin, Ireland. Association for Computational Linguistics.

[9] 松下佳世，山田優，石塚浩之，“英日・日英通訳データベース（JNPC コーパス）の概要”，日本翻訳学会，vol. 22, pp. 87–94, 2020

A 付録

MQM による同時通訳評価者と翻訳評価者の同時通訳データ評価結果比較
データ：JNPC063

カテゴリ	原言語	目的言語	同時通訳者評価（エラー値）	翻訳評価者（エラー値）
正確さ-誤訳-	and they also need to be well qualified to take on the duties that will inevitably fall on his shoulders	そしてしっかりとした資格を持ってやはり彼らの肩にかかってくる今後の義務を担ってもらわねばならない。	なし(0)	「well qualified」を「資格を持って」と訳すのは不適切(1.33)
正確さ -抜けと余分-	It's very sad that I discovered before before the elections in 2015 I was surveying the situation in my country	私も 2015 年の選挙の前に分かったことですけれども状況の調査をしておりました。	"It's very sad" that ～が（抜け）→「非常に残念なことで～」(2.66)	「It's very sad that I discovered 」が訳出されていない。(1.33)
正確さ -抜けと余分-	We have never had the opportunity they have never had the opportunity in my country	彼らには機会が与えられてこなかったからであります。	なし(0)	in my country が訳されていない(1.33)
正確さ -抜けと余分-	it may be a surprise to some of you to learn that we haven't had proper campus life at the universities since well since about 1962.	皆さん 驚く かもしれません が ミャンマー では 大学 でも 適切な 十分な キャンパス ライフ は 送れ なかった	”since about 1962" が抜け。「1962 年以降ずっと」が正しい(2.66)	since about 1962. が未翻訳(1.33)

重みの詳細

エラー・カテゴリ設定			重大度設定		重み付け設定	
カテゴリ一覧	**重み設定**	**相対重み**	**重み名**	**点数**	**重み名**	**点数**
正確さ	とても重視	1.33	とても重視	2.0	とても重視	2.0
正確さ-誤訳	とても重視	1.33	やや重視	1.5	やや重視	1.5
正確さ-抜けと余分	とても重視	1.33	普通	1.0	普通	1.0
正確さ-未翻訳	とても重視	1.33	あまり重視しない	0.5	あまり重視しない	0.5
正確さ-その他	とても重視	1.33				
流暢さ	普通	0.67				
流暢さ-文法誤り	普通	0.67				
流暢さ-誤用	普通	0.67				
流暢さ-その他	普通	0.67				